A guided approach for automatic segmentation and modeling of the vocal tract in MRI images

Abdulkadir Eryildirim, Marie-Odile Berger

A GUIDED APPROACH FOR AUTOMATIC SEGMENTATION AND MODELING OF THE VOCAL TRACT IN MRI IMAGES

Abdulkadir Eryildirim, Marie-Odile Berger

LORIA/INRIA Nancy Grand Est, France
email: Abdulkadir.Eryildirim@loria.fr,Marie-Odile.Berger@loria.fr

ABSTRACT

An automatic approach for segmenting tongue contours in physical correspondences is described in this paper. The goal is to be able to segment large MRI database and to automatically produce vocal tract models from these data. We extend the guided method presented in [6] in order to obtain physically corresponding curves and provide in-deep evaluation of the method.

1. INTRODUCTION

Articulatory modeling of the vocal tract, and especially the tongue, is crucial for many applications. Speech training for hearing impaired children or second language learning is one example, where the visual feedback can efficiently supplement the auditory feedback.

Magnetic resonance imaging (MRI) is a powerful tool in obtaining the vocal-tract geometry and does not involve any known radiation risk. For the vocal tract shaping to be studied in speech production, many images are acquired from a particular subject for different utterances of interest and must be segmented. Delineation of vocal tract can be performed manually with great accuracy but this task is labor-intensive for large data sets [7]. Methods for automatic segmentation are thus requested but this is a difficult issue for several reasons. First, the tongue is the most flexible organ of all the active articulators. It could move near other edges in the oral cavity, such as the palate, the lips and the teeth, which may disturb the segmentation process. Second, due to a quite long acquisition time of MRI data, the speaker is required to artificially sustain a sound which may induce motion blur. Bresh and Narayanan [1] have recently proposed a method for unsupervised region segmentation using its spatial frequency domain representation. However the method requires considerable supervision to initialize and do not really take into account anatomical constraints on the vocal tract. On the contrary, we have proposed in [6] a method which integrates shape priors learned on a reference speaker to guide the segmentation of a new speaker. Though the method has proven to be efficient for segmenting the tongue in preliminary experiments, building a model of the vocal tract from the output of the algorithm is not straightforward since the curves which are obtained are not in physical correspondences. This prevents us to use the widely used PCA approach [3] to build a model. The goal of the current study is to improve the model presented in [6] in several ways, especially to obtain segmented tongues in physical correspondences. The main ideas of [6] are briefly reviewed in section 2.

2. A VARIATIONAL FRAMEWORK FOR TONGUE SEGMENTATION

The idea explored in [6] is to guide the segmentation with shape priors learned on a reference speaker within a shape-based variational framework. Shape priors are incorporated into segmentation via a PCA model with a relatively large number of components to enable the adaptation of the model to strong morphological differences of new speakers. A PCA model is build from manually delineated contours of the tongue of a reference speaker. As a result, when using the p eigenvectors corresponding to the largest eigenvalues, a new shape, C, of the same class as the training set, can be approximated by $C(w) = \bar{C} + \sum_{i=1}^{p} w_i \delta C_i$, where $\bar{C}$ is the mean shape, and δC_i is the eigenvector of the covariance matrix of the data. Choosing p is crucial. Often p is chosen so as to explain a given proportion of the variance exhibited in the training set (97%, here 6 modes). In our case, the model is based on a reference speaker but is supposed to adapt to new speakers. We thus go beyond this value and considered in the initial article 15 modes (99%) to allow for generalization to new speakers.

Given this curve model, an energy functional is defined for segmenting the tongue. It is composed of two region-based energies:

$$E = \alpha E_G + E_L, \tag{1}$$

where the *global* energy term E_G describes image information about pixels in the whole image; while the *local* energy term E_L introduces image information inside a small neighborhood of points along the contour C. The global energy term E_G adopts the Chan-Vese model proposed in [2]. E_G computes the optimal approximation of an image I as a piecewise constant binary function. It is written as

$$E_G(C) = \int_{C_{\text{in}}} \big(I(\mathbf{x}) - \mu\big)^2 d\mathbf{x} + \beta \int_{C_{\text{out}}} \big(I(\mathbf{x}) - \nu\big)^2 d\mathbf{x}$$

Each curve is represented by a set of equidistant points. A gradient descent method is used to minimize E with respect to the w eigencoefficients. The initialization curve is the tongue contour obtained with the reference speaker for the same sound. In order to guaranty plausible shapes, w_i coefficients are constrained to belong to the interval $[-flex\sqrt{\lambda_i}, flex\sqrt{\lambda_i}]$, where λ_i are the eigenvalues of the covariance matrix of the training set. Whereas $flex = 3$ is classically used, we here use larger values ($flex = 5$) to enhance the generalization capabilities of the model to new speakers.

With respect to this model, the contribution of this paper is two-fold:

- We experimentally show that the reference model is able to cope with strong morphological differences between

speakers with a limited numbers of modes. Though important, this validation was not addressed in [6].

- We propose an automatic method for the identification of the end points as well as an improved variational framework to obtain curves in physical correspondences.

3. VALIDATION OF THE REFERENCE MODEL

The goal of this section is to show that the extended model we use

$$C(w) = \bar{C} + \sum_{i=1}^{p} w_i \delta C_i \;\; with\; w_i \in [-flex\sqrt{\lambda_i}, flex\sqrt{\lambda_i}]$$

has sufficient generalisation capabilities to handle speakers with important morphological differences and brings sufficient dimensionality reduction and shape priors to be used in the optimization framework. A reference speaker, which has the most MRI volumes (38) designed so as to cover the range of French articulation as much as possible, and three test speakers with high morphomological differences - see Fig. 2- were considered in this study : two male speakers named M_1 (27 sounds) and M_2 (14 sounds) and a female speaker F_1 (9 sounds). For each speaker, 3D/3D affine registration based on mutual information was performed to compensate for different head positions between acquisitions. Affine registration was performed between the different speakers to allow model adaptation. We only consider in this paper 2D models built in the midsagittal plane after intra and inter speaker registration. The image size is 512x512 and the pixel size is .625 mm.

Since the origin of the coordinate system of the test speaker is not known, a translation vector T is added as a supplementary parameter to the main PCA equation to adjust the origin. The new description of shapes is then given by:

$$C(w) = T + \bar{C} + \sum_{i=1}^{p} w_i \delta C_i \qquad (2)$$

Our goal is to determine the appropriate number of PCA components and the appropriate $flex$ value , which are convenient for representing new speakers with strong morphological differences. We want to find the smallest p and the appropriate $flex$ that are able to represent these speakers in order to avoid that extra parameters give unwanted variabilities to the model when submitted to noise contained in the image during the segmentation process. Another reason is that modes with higher ranks tend to explain noise rather than variabilities and may be not relevant in the linear basis. It is important to note here that the value of p has nothing to do with the number of PCA components that are used to explain a given speaker (generally 6-8 components). Here higher values of p are used to generalize the reference model to the shape space of new speakers.

Determination of appropriate values of p and $flex$ is done by computing, for several p and $flex$ values, the curve produced by the model which best fits the ground truth, that is the curve manually detected by an expert. Distances between the ground truth and the curve produced by the model for different values of p and $flex$ are shown in table 1. The curves were resampled with n (n=40) equally spaced points and the distance is the mean of the distance between points of same index. Table 1 contains the average of these distances over all sounds produced by the three speakers.

Table 1: Generalization performance of the PCA model

No of modes	flex	Stat. #1	Stat. #2	Stat. #3	Stat. #4
5	5	%0.00	%2.04	%10.20	%24.48
5	3	%0.00	%0.00	%8.16	%22.44
10	5	%8.16	%34.69	%75.51	%100
10	3	%8.16	%30.61	%40.82	%75.51
15	5	%36.73	%81.63	%97.96	%100
15	3	%26.53	%51.02	%63.27	%89.80
20	5	%63.27	%95.92	%100	%100
20	3	%34.70	%53.06	%65.31	%89.80
25	5	%91.84	%95.92	%100	%100
25	3	%40.82	%57.14	%65.31	%89.80

Table 1 exhibits four statistics named Statistics 1,2, 3 and 4 which give the percentage of sounds which provides mean Euclidean distance lower than 0.625mm (1 pixel), 0.9375mm (1.5 pixels), 1.25mm (2 pixels) and 1.875mm (3 pixel) respectively.

In general, as expected and also confirmed by Statistics 1, 2, 3 and 4 in Table 1, increasing number of components also provides better fitting measures. Furthermore, from Table 1, it is concluded that using larger PCA constraint, $flex = 5$, provides better fitting. With $flex = 5$, for all sounds, Mean Euclidean distance lower than 1.875mm are achieved for only 10 PCA modes and making $flex = 3$ significantly decreases the values of Statistics 1, 2 and 3. As seen from the plot of the articulation 'u' of M_1 which gives unsatisfactory performance in terms of the measures in Figure 1, using 5 for PCA constraint $flex$ can decrease the distance measure by half and more.

In conclusion, for general coverage, with 15 PCA modes, the PCA model achieves fitting to the %97.96 of the articulations with mean distance lower than 1.25mm. The only exception is the articulation 'u' of M_1, whose fitting measures are shown in in Figure 1. With 20 parameters and $flex = 5$, the PCA model is able to generalize the three speakers with %100 as seen in Table 1. Consequently, with no more than 20 parameters, the PCA model of the reference exhibits sufficient generalisation capabilities to handle speakers with important morphological differences.

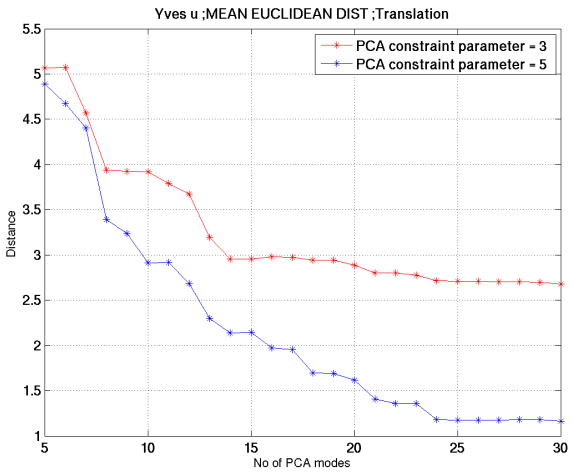


Figure 1: Fitting the PCA model to the sound 'u' of M_1.

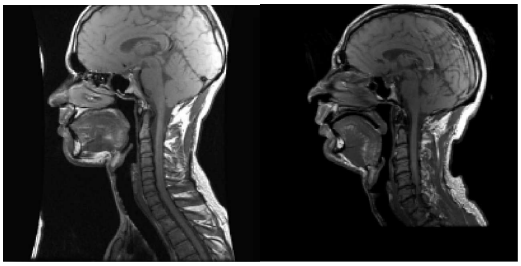


Figure 2: Morphological differences betwen the reference and the M_1 speaker.

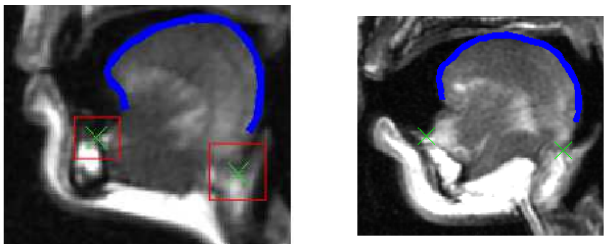


Figure 3: Segmentation results of the procedure described in [6] for F1 (o) and M2 (ka). Green crosses indicates the extremities of tongue manually labelled. Correlation windows are drawn in red.

4. FROM SEGMENTATION TO TONGUE MATCHING

4.1 Automatic detection of end points

One of the main drawback of the method in [6] is the lack of information about the extremities of the tongue. As seen in Fig. 3, segmentation does not provide curves with physically corresponding extremities. In order to incorporate physical constraints on the extremities of tongue into the segmentation process as well as to eliminate need for human intervention during segmentation, an automatic method which aims to detect the end points is proposed. The method is based on a two-step template matching approach for detecting the bottom of the lingual frenulum and the end of the epiglottis for all sounds. These points are displayed with green crosses in Fig. 3.

Since the shape and the photometry around these points dramatically change according to the speaker, we proceed by providing the system with one template of the two points for one sound for each speaker. Examples of templates are shown in red in Fig. 3. The system then automatically finds the location of the template for the other sounds of the same speaker using normalized cross-correlation. The resulting point is only an approximation of the expected point since correlation only produces the translation of the pattern that best fits the current sound. Table 2 summarizes the average errors in pixels for each speaker between manually detected points and the points obtained with the automatic correlation based method. Examples of detections are provided in Fig. 4.

A way to improve the detection is to take into account the fact that the pattern, especially around the epiglottis, is submitted to elastic deformation whereas we only consider linear transformations in our procedure. We thus use non rigid registration methods [4] in order to match the pattern predicted with correlation and the current sound. This allows us to refine the position of the extreme points: the new position is the image of the last position under the elastic transformation. First experiments provided superior results (see table 2. With speaker F_1, this additional step decreased average error between detected and ground-truth end points of the tongue from 12.03 pixels to 7.21 pixels.

Table 2: Average errors for automatic extreme point detection based on correlation scheme

Speaker	Average Error For Front Extremity	Average Error For End Extremity
F_1	2.5	12.03
M_1	5.82	5.84
M_2	6.47	8.44

4.2 Incorporating constraints into the variational framework

In order to make use of the information of extremities extracted using automatic detection, a new term E_{EXT} is included into the existing energy functional E such that:

$$E = \alpha E_G + E_L + E_{EXT} = E_{old} + E_{EXT} \quad (3)$$

where E_{EXT} is an energy functional whose minimization drives the extremities of evolving curve towards the actual extremities. E_{EXT} is defined as :

$$E_{EXT} = \kappa_1 E_{front} + \kappa_2 E_{end} \quad (4)$$

E_{front} (resp E_{end}) defines the cost to the fitting to the actual front (resp end) tip of tongue. κ_1 and κ_2 are used to determine the degree of fit to the extreme points as well as the contribution of the new functional over the curve evolution. Let a and b be the end points. E_{front} and E_{end} penalize the deviations from the given extremities and are defined as:

$$E_{front} = (a - T - \overline{C}(0) - \sum_{i=1}^{p} w_i v_i(0))^2 \quad (5)$$

$$E_{end} = (b - T - \overline{C}(n) - \sum_{i=1}^{p} w_i v_i(n))^2 \quad (6)$$

where n is to the number of points of the contour.

Minimization of our new functional with respect to PCA model parameters w leads to the equations:

$$\frac{\partial E_{front}}{\partial w_i} = -2v_i(0)(a - T - \overline{C}(0) - \sum_{i=1}^{p} w_i v_i(0)) \quad (7)$$

$$\frac{\partial E_{end}}{\partial w_i} = -2v_i(n)(a - T - \overline{C}(n) - \sum_{i=1}^{p} w_i v_i(n)) \quad (8)$$

The updated curve evolution equation becomes :

$$\frac{\partial w_i}{\partial t} = -\frac{\partial E_{old}}{\partial C} \cdot v_i - \frac{\partial E_{front}}{\partial w_i} - \frac{\partial E_{end}}{\partial w_i}, \forall i \in \{1, ..., p\} \quad (9)$$

In order to equally balance both extremities of tongue, equal weights are used ($\kappa_1 = \kappa_2$).

This updated energy functional is used in the second step of a two-step procedure. In the first step, the aim is to reach the best global boundaries using the previous functional E_{old} (Eq. 1). In the second step, the set of extreme points are used and the new energy (3) is minimized from the curve obtained at the first step allowing us to obtain curves in physical correspondences.

5. EXPERIMENTAL RESULTS

Results of the complete segmentation process are presented in this section. Our method is quantitatively evaluated using the difference computed between the automatically detected contour and the ground truth drawn by an expert. In order to be independent from curve discretization, we use a Mean Sum of Distances (MSD) which measures the distance between the closest contour elements of each curve: Let $U = [u_1,..,u_n]$ and $V = [v_1,..,v_n]$ be two discretized curves, the MSD distance is defined as $MSD(U,V) = \frac{1}{2n}(\sum_{i=1}^{n} min_j d_{euc}(v_i,u_j) + \sum_{i=1}^{n} min_j d_{euc}(u_i,v_j))$, where d_{euc} is the Euclidean distance between points.

Due to morphological differences, parameters of the algorithm are adjusted for each speaker. By visual inspection, it is observed that $flex = 3$ is better for F_1 and M_2 while $flex = 5$ is suitable for M_1. Consequently, the PCA constraints are set to these default values.

Increasing the values of κ in the new functional makes the distance between the extreme points given and the extreme points of the evolving curve closer. This distance was calculated for different values of κ. Based on our experiments, for lower κ values, the Euclidean distance drops sharply but after $\kappa = 2000$, it decreases very slowly. Therefore, $\kappa = 2000$ is used in the segmentation.

Segmentation results are given in table 3 which displays the MSD distance between ground truth curves and the curves obtained with our method for the M_1 speaker (28 sounds). Five statistics named 1, 2, 3, 4 and 5 are displayed. They give the percentage of sounds which provides MSD distance lower than 1.25mm (2 pixels),1.56 mm (2.5 pixels), 1.875 mm (3 pixels), 2.18mm(3.5 pixels) and 2.5mm (4 pixels) respectively. Three methods are considered:

- Initial: Method decribed in [6] which involves only E_G and E_L (first step)
- Auto : refers to the proposed procedure, which uses automatically detected end points into E_{EXT}.
- Manual : refers to the procedure which uses manually detected end points into the functional E_{EXT}

From this table, it is clear that our new method which imposes physical correspondences outperforms the initial method. For 20 modes, the mean MSD errors in pixels are respectively 3.9 (Initial), 2.57 (Auto), 2.33 (Manual). These results have to be compared to the delineation errors made by the expert when several delineations are realized at distant times. Due to motion blur in the MRI images, accurate manual detection is difficult and prone to error and leads to a MSD distance of 2.06 pixels between delineations realized by our expert at two distant times. With a mean error of 2.57 pixel for the automatic method, we are thus close to the delineation error.

Table 4 provides results with M_2 speaker who have large morphomological differences with the reference speaker. Here again, the segmentation is dramatically improved with our method.

Figure 4 illustrates the segmentation results for the method in [6] and for the method described in this paper and referred as 'Auto'. From visual inspection, it is obvious that our method provides superior fitting to the extremities while preserving the smoothness and plausibility.

Fig. 5 exhibits the curves obtained for sounds 'li' and 'o' with the automatic method as well as the curve delineated by the expert. These curves are visually in good agreement with the images and the distances with ground truth are respectively 2.45 and 1.73 pixels.

Table 3: Segmentation Results for M_1 speaker.

Nb of modes	Method	Stat. #1	Stat. #2	Stat. #3	Stat. #4	Stat. #5
10	Initial	%0.0	%0.0	%15.4	%23.1	%50.0
10	Auto	%0.0	%0.0	%38.46	%65.4	%80.76
10	Manual	%0.0	%19.23	% 46.15	% 73.1	% 76.9
15	Initial	%0.0	%0.0	%19.2	%38.4	%57.6
15	Auto	%3.8	%42.3	%76.9	%82.3	%100
15	Manual	%26.9	%73.1	% 84.6	% 100	% 100
20	Initial	%0.0	% 7.7	%19.2	%30.7	%65.38
20	Auto	%11.5	%42.3	%80.8	%100	%100
20	Manual	%26.9	%73.1	% 88.4	%100	%100

Table 4: Segmentation results for M_2 speaker.

Nb of modes	Method	Stat. #1	Stat. #2	Stat. #3	Stat. #4	Stat. #5
20	Initial	%0.0	% 21.4	%28.5	%50.0	%71.4
20	Auto	%7.14	%35.7	%71.4	%85.7	%92.8
20	Manual	%28.5	%64.2	% 78.5	%92.8	%100

In order to observe the effect of the additional elastic registration step in the automatic detection procedure, Table 5 compares results obtained with the classical correlation process, referred as 'Corr' and results obtained with the additional elastic registration step, referred as 'Elas' on F_1. The statistics displayed are the same as in table 3 and show that the use of elastic correlation noticeably improves the accuracy of the segmentation process.

Finally, an important way to assess the results is to compare the PCA modes computed for one speaker when manual and automatic segmentation are used. Results are here provided for speaker M_1 with 27 sounds. The modes we obtain are very close. The first seven eigenvalues which explain more than 97% of the data are $[56.2, 28.3, 23.0, 19.95, 12.96, 10.36, 6.5]$ for manual detection and $[52.8, 29.5, 23.1, 19.96, 13.8, 8.99, 7.44]$ for auto-

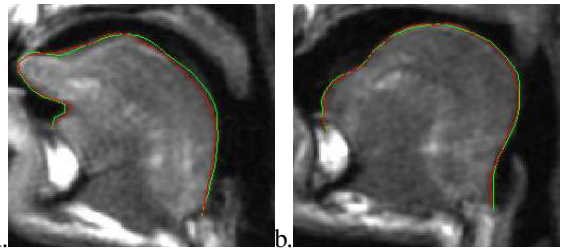
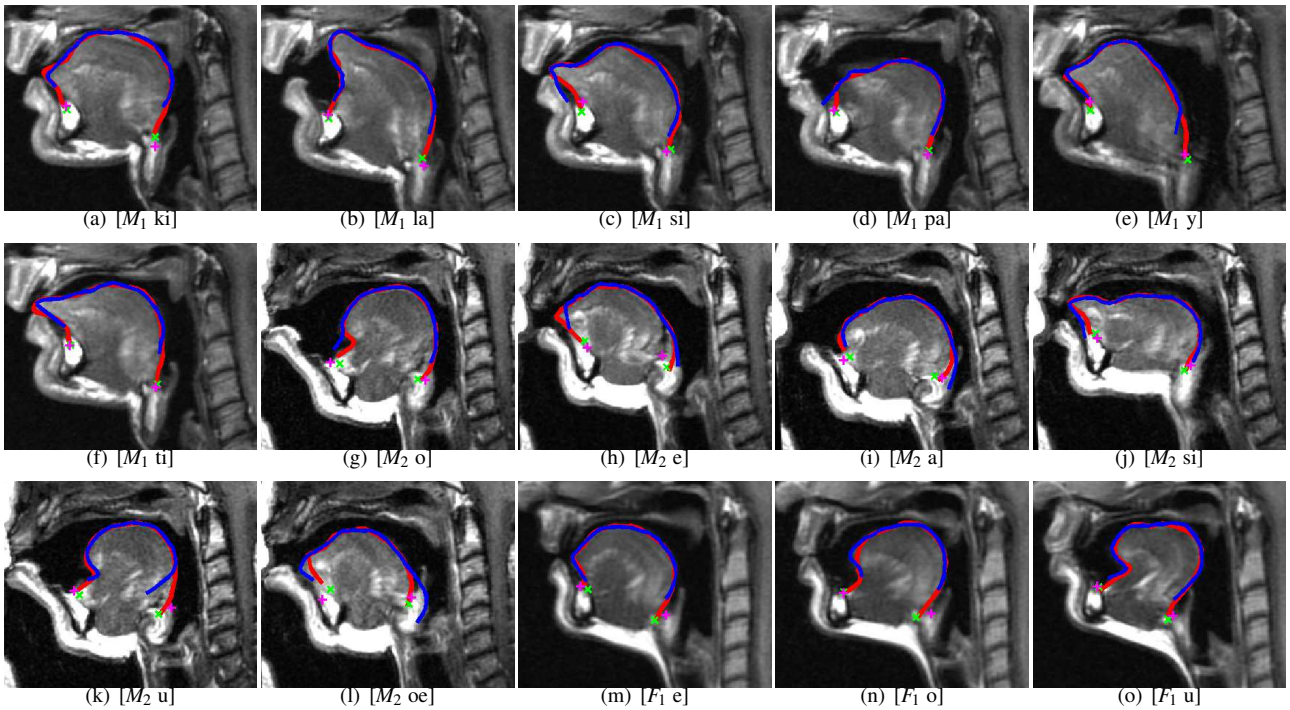


Figure 5: Ground truth (green) and automatic segmentation (red) on M_1. (a) for *li*, mean distance is 2.45 pixel (1.53 mm) , (b) for *o* mean distance is 1.73 pix (1.08mm).



(a) [M_1 ki] (b) [M_1 la] (c) [M_1 si] (d) [M_1 pa] (e) [M_1 y]

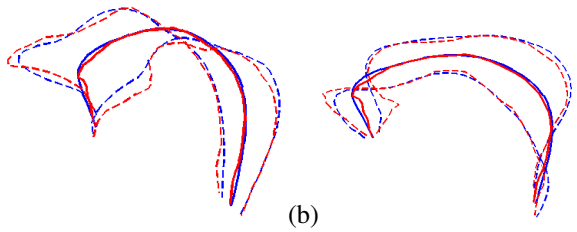
(f) [M_1 ti] (g) [M_2 o] (h) [M_2 e] (i) [M_2 a] (j) [M_2 si]

(k) [M_2 u] (l) [M_2 oe] (m) [F_1 e] (n) [F_1 o] (o) [F_1 u]

Figure 4: Segmentation Results on the mid-sagittal images of M_1, M_2 and F_1. Blue curve: the method in [6]; red curve: final result obtained using our new automatic segmentation procedure (Auto) proposed in this paper. Green crosses are automatically detected extremities using correlation, pink plus signs are manual extremities

Table 5: Segmentation results with elastic correlation for F_1

No of modes	Method	Perc. #1	Stat. #2	Stat. #3	Stat. #4	Stat 5
15	Corr	%0.0	%25.	%37.5	%62.5	%75
15	Elas	%0.0	%0.0	% 75.0	% 75.0	%87.50
20	Corr	%0.00	%25.00	%37.5	%75	%87.5
20	Elas	%0.0	%12.5	%62.5	%87.50	% 87.5



(a) (b)

Figure 6: The first (a) and second (b) modes obtained with manual delination (blue) and automatic segmentation (red).

matic detection. Fig. 6 displays for the first two modes the mean curve (in bold) and the curves $mean + 3\sigma_i mode_i$ and $mean - 3\sigma_i mode_i$ (dotted lines). Modes obtained with manually delineated data are drawn in blue and those obtained with the automatic segmentation process are in red. The results are very similar, proving the reliability of our method.

6. CONCLUSION

We have presented and evaluated a method for automatic detection of tongue contours in physical correspondences. Results are really promising and we intend to use this method to automatically extract contours in larger MRI database which will be acquired in our laboratory in the near future. These curves in correspondence are of particular interest since they allow automatic construction of a PCA model for a particular speaker. They can also be used to study how model adaptation can be performed between speakers.

REFERENCES

[1] E. Bresch and S. Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *ITMI*, 28(3), 2009.

[2] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Trans. on Image Processing*, 10(2):266–277, 2001.

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *PAMI*, 23(6):681–684, June 2001.

[4] Dirk-Jan Kroon and Cornelis H. Slump. Mri modality transformation in demon registration. In *Int. Symposium on Biomedical Imaging:*, pages 963–966., 2009.

[5] C. Li, C. Kao, J. Gore, and Z. Ding. Implicit active contours driven by local binary fitting energy. In *Computer Vision and Pattern Recognition*,2007.

[6] Ting. Peng, Erwan Kerrien, and Marie-Odile Berger. A shape base framework to segmentation of tongue contours from MRI data. In *ICCASP 2010*, 2010.

[7] M. Stone, Davis, Douglas A. E., M. Aiver, Gullapalli R., W. Levine, and A. Lundberg. Modeling tongue surface contours from cine-mri images. *Journal of Speech, Language, and Hearing Research*, 44:1026–1040, 2001.